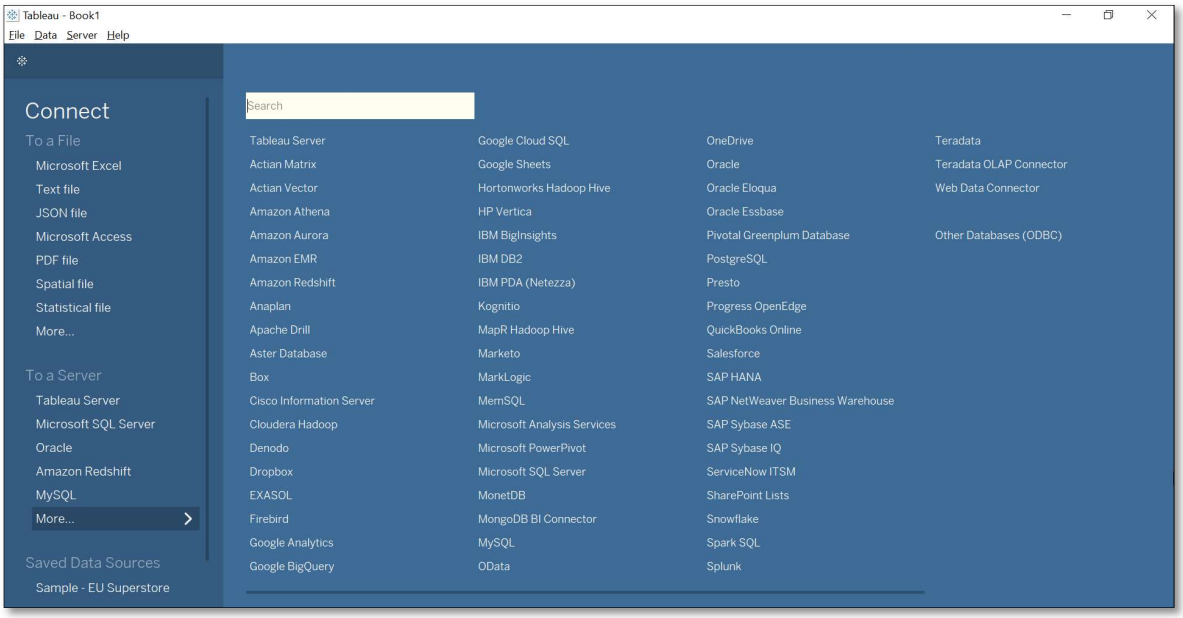
Module 1 – Connecting to Data

Contents:

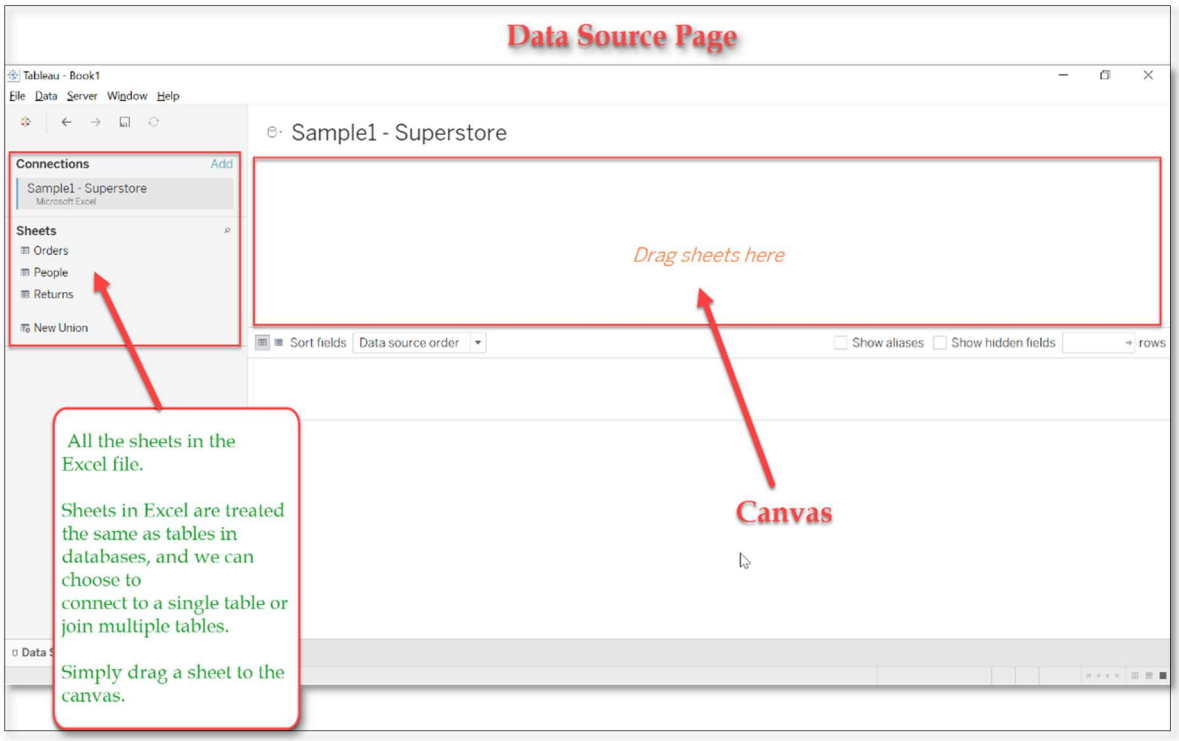
1. Connecting to Data

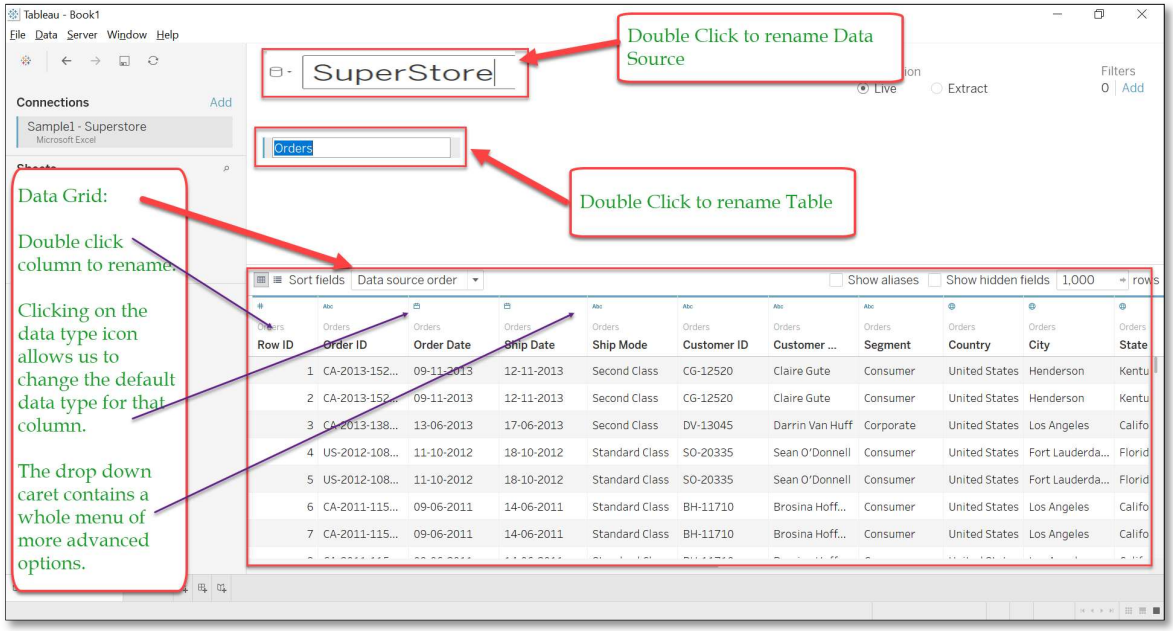
a. Connecting to an excel file

Tableau can connect to many data sources. In the Connect pane, we have a long list of native connections to all sorts of data sources, regardless of whether they're local files or databases on-premises or in the cloud. Example: Global Superstore Excel. Superstore is a data set of sales for a global retail chain that sells furniture, office supplies, and technology goods. Each row of data represents a single item in a transaction.

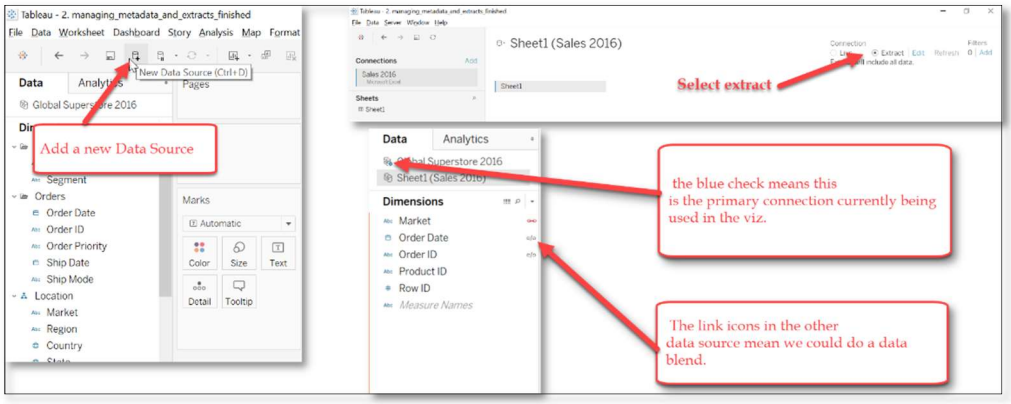


b. The data source page

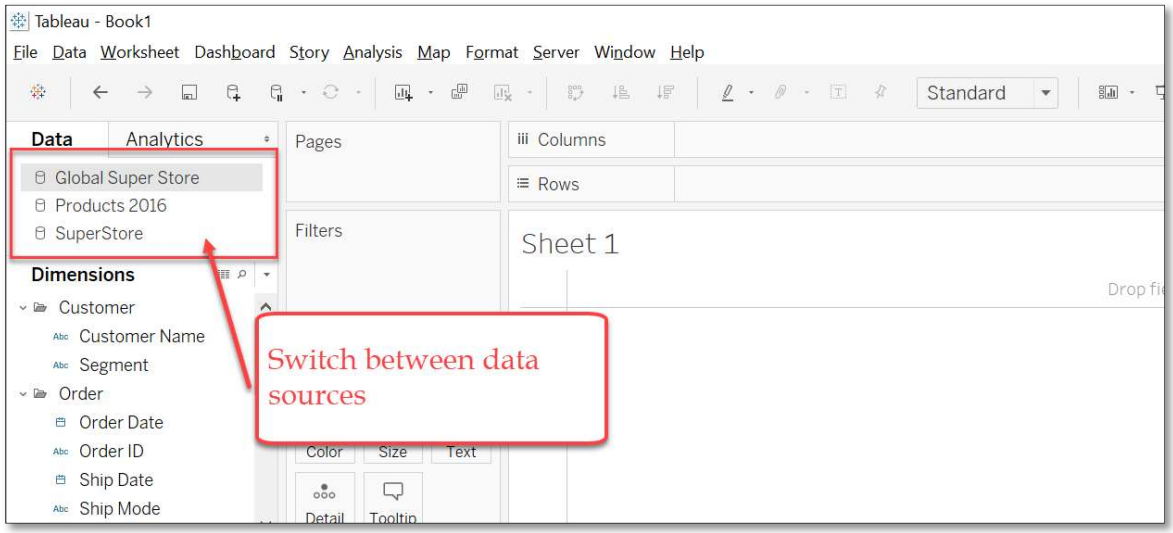




c. Changing Data Sources

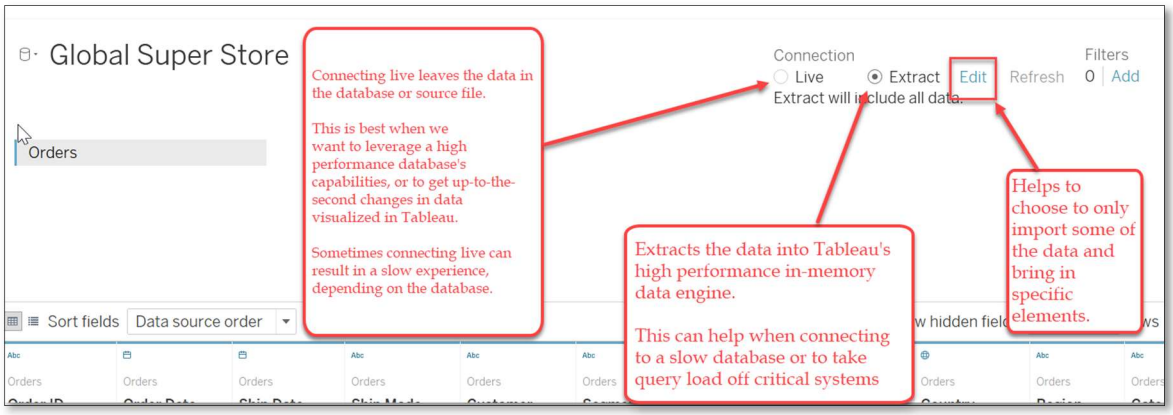


d. Switching between multiple Data Sources

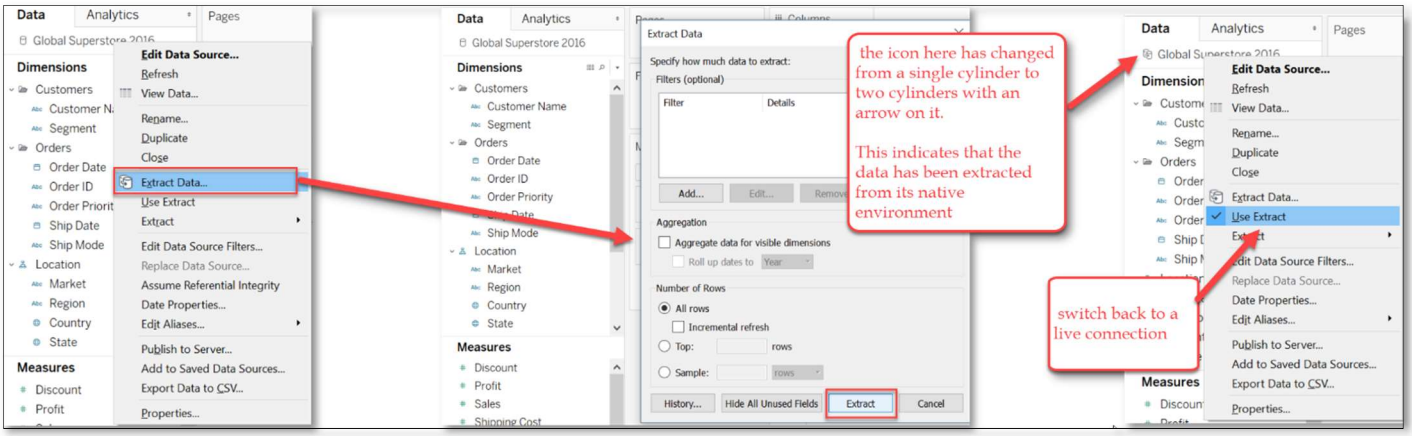


2. Managing Extracts

a. Live versus extract



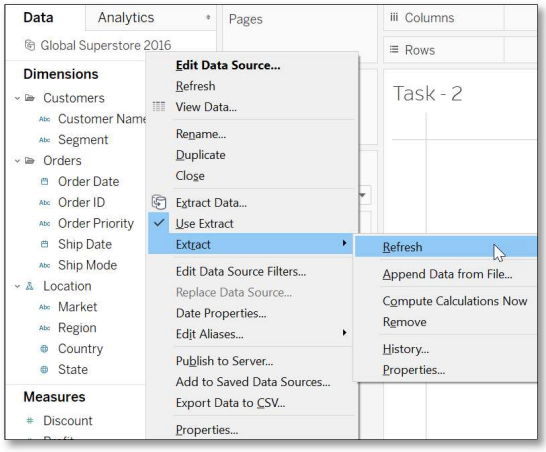
b. Creating Extracts



If the workbook has already been saved as a .twbx packaged workbook, the extract will automatically be saved as part of the packaged workbook. If the workbook has been saved as a .twb or hasn't been saved yet, we'd be asked where to save the Tableau Data Extract file (.tde) or as (.hyper).

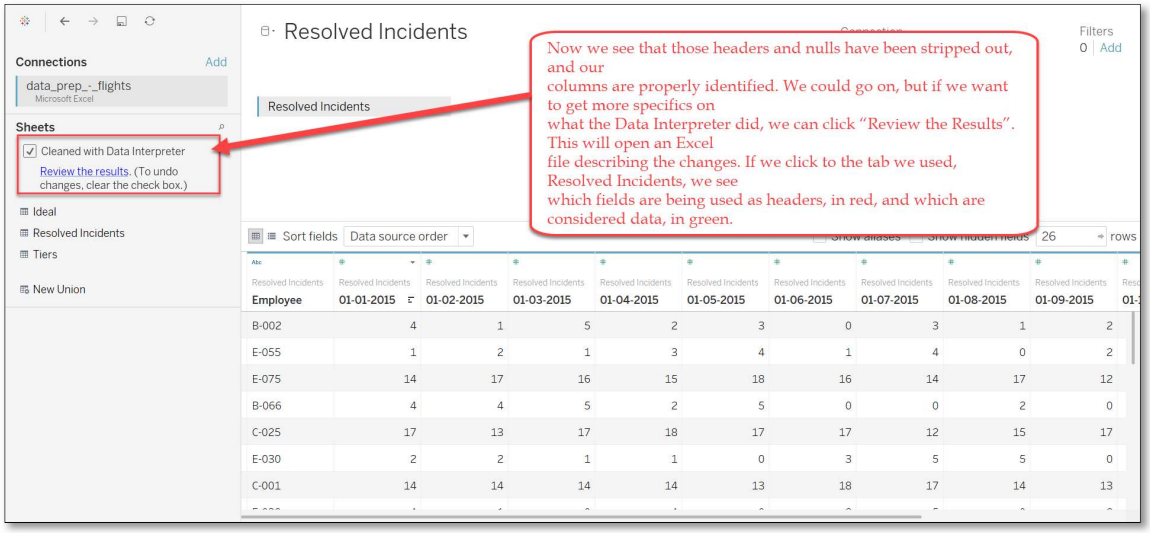
c. Refreshing Extracts

Because we're currently connected to a static copy of the data, if the underlying data were to change, the view would not update with that new data until the extract is refreshed.



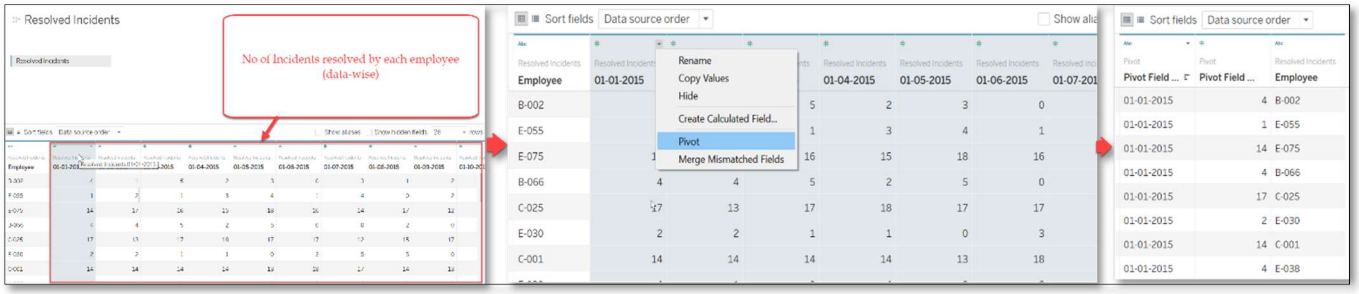
3. Data Prep with Text and Excel Files

a. Data interpreter



b. Pivot

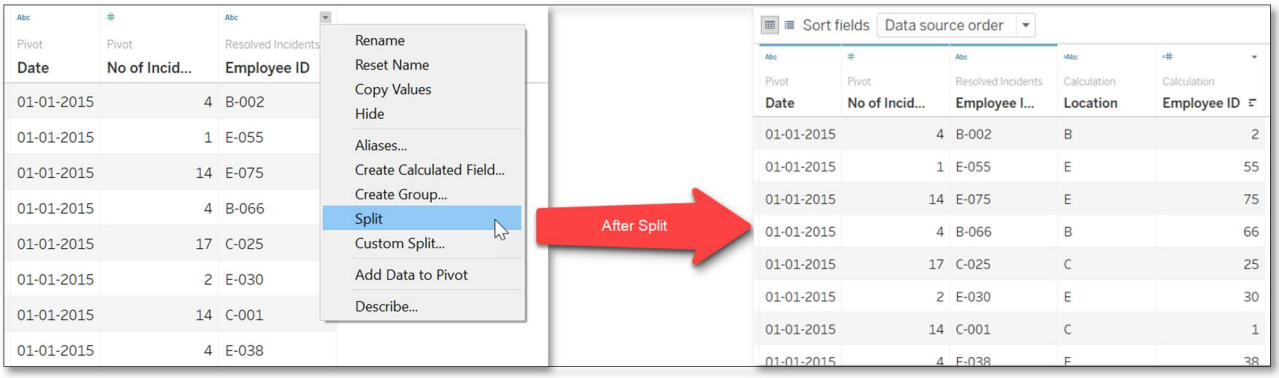
To change the format from that column-per-month layout into a single date column and a single column for Resolved Incidents, select all the date columns, open the menu and select "Pivot". This pivot feature essentially merges the information from the original columns and rows into two new columns – Pivot field names, and Pivot field values. Rename the fields to suitable names.



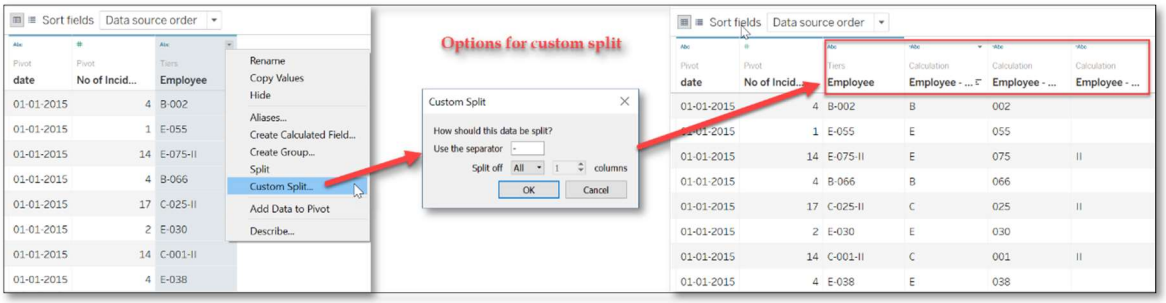
c. Pivot using custom SQL

d. Metadata grid

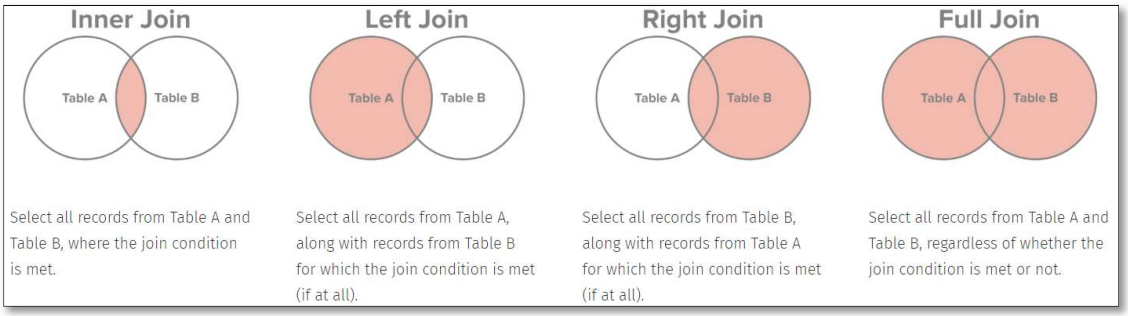
e. Split



f. Custom split



4. Join Types with Union



Example:

Customer Table:

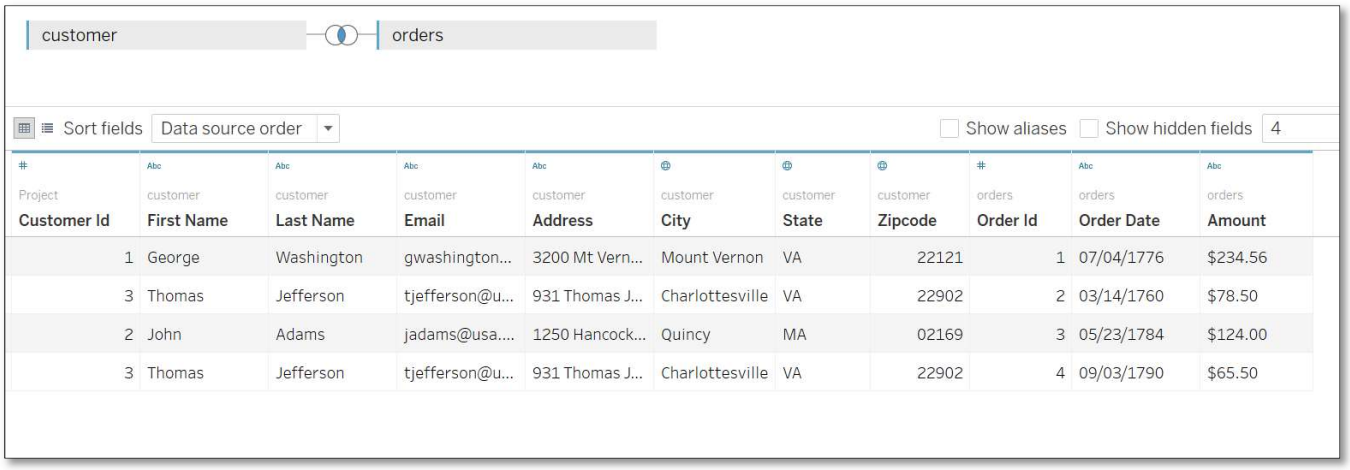
customer_id	first_name	last_name	email	address	city	state	zipcode
1	George	Washington	gwashington@usa.gov	3200 Mt Vernon Hwy	Mount Vernon	VA	22121
2	John	Adams	jadams@usa.gov	1250 Hancock St	Quincy	MA	2169
3	Thomas	Jefferson	tjefferson@usa.gov	931 Thomas Jefferson Pkwy	Charlottesville	VA	22902
4	James	Madison	jmadison@usa.gov	11350 Constitution Hwy	Orange	VA	22960
5	James	Monroe	jmonroe@usa.gov	2050 James Monroe Parkway	Charlottesville	VA	22902

Orders Table:

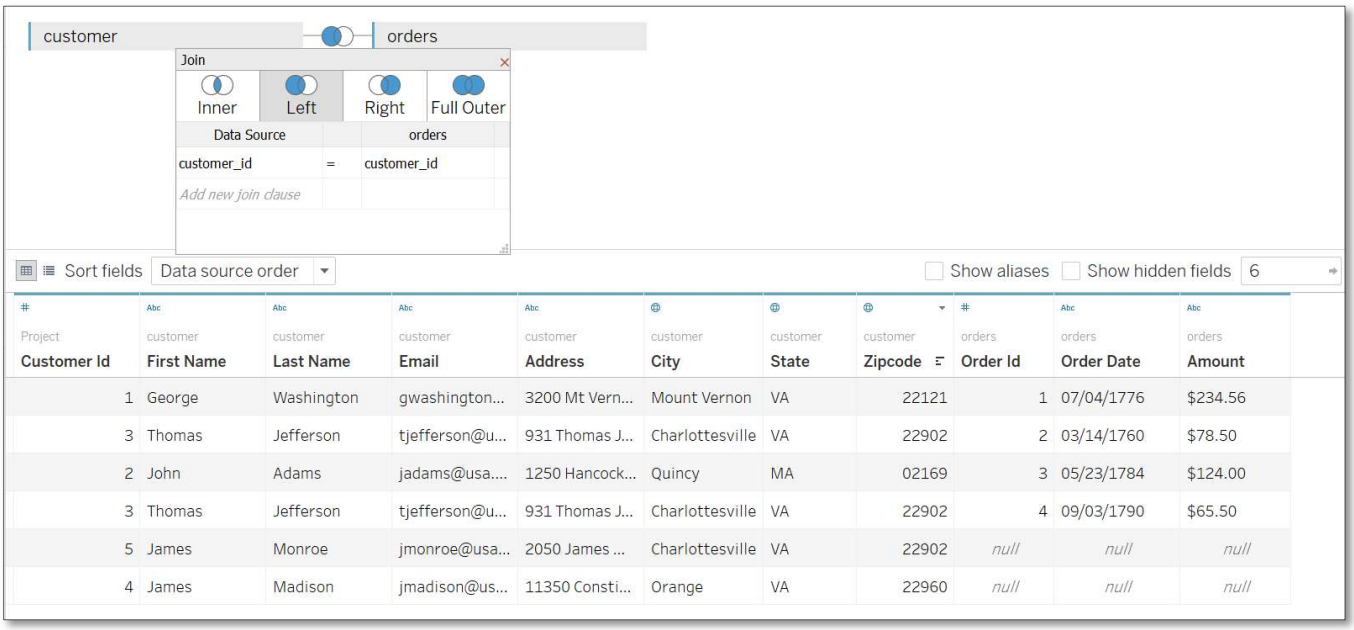
order_id	order_date	amount	customer_id
1	07/04/1776	$234.56	1
2	03/14/1760	$78.50	3
3	05/23/1784	$124.00	2
4	09/03/1790	$65.50	3
5	07/21/1795	$25.50	10
6	11/27/1787	$14.40	9

Note that (1) not every customer in our customers table has placed an order and (2) there are a few orders for which no customer record exists in our customers table.

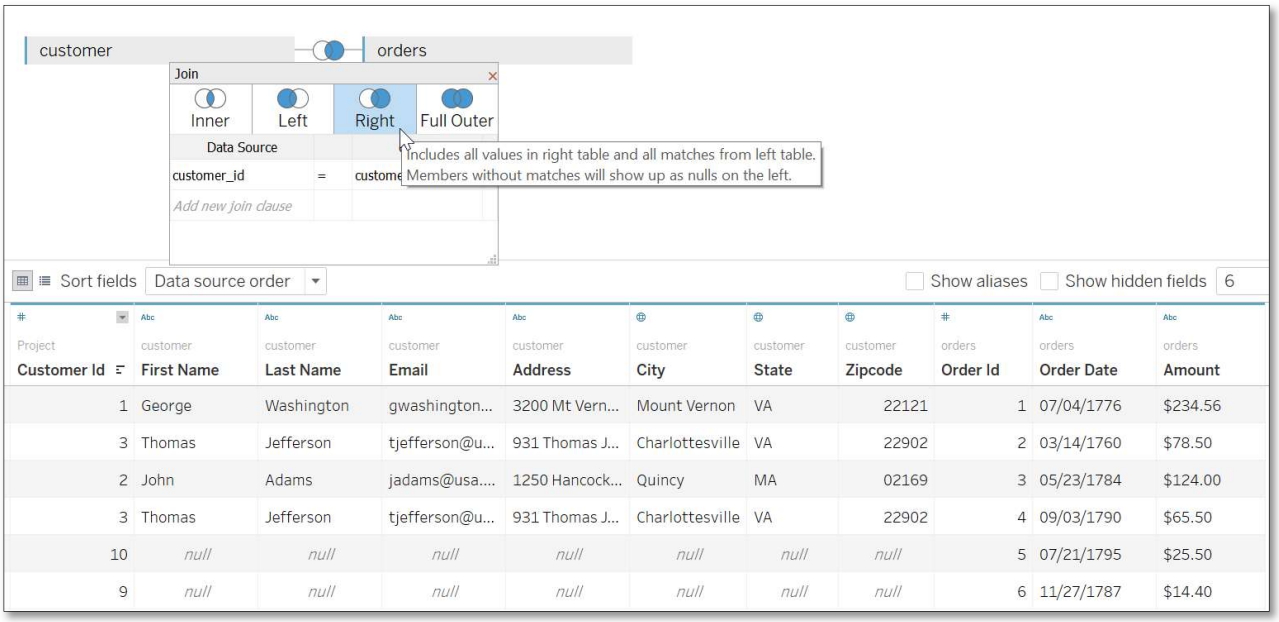
Inner Join



Left Join

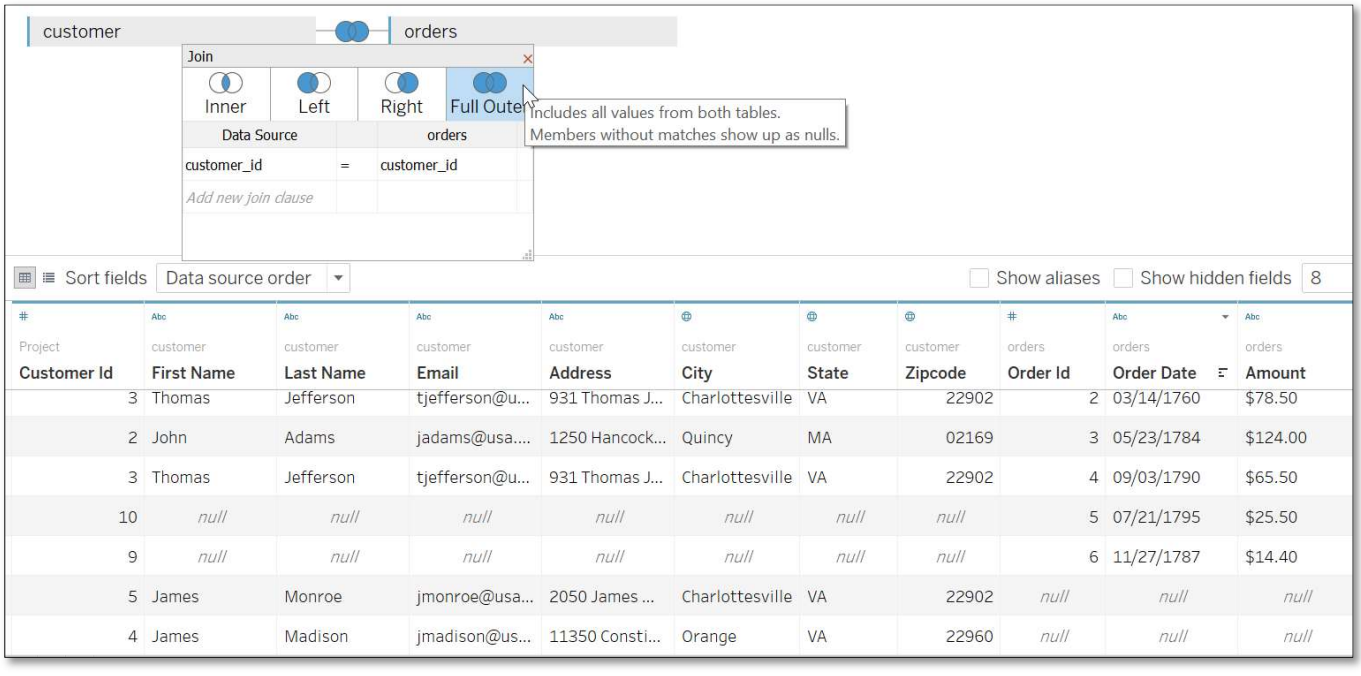


Right Join



Project Customer Id	customer First Name	customer Last Name	customer Email	customer Address	customer City	customer State	customer Zipcode	orders Order Id	orders Order Date	orders Amount
1	George	Washington	gwashington...	3200 Mt Vern...	Mount Vernon	VA	22121	1	07/04/1776	$234.56
3	Thomas	Jefferson	tjefferson@u...	931 Thomas J...	Charlottesville	VA	22902	2	03/14/1760	$78.50
2	John	Adams	jadams@usa....	1250 Hancock...	Quincy	MA	02169	3	05/23/1784	$124.00
3	Thomas	Jefferson	tjefferson@u...	931 Thomas J...	Charlottesville	VA	22902	4	09/03/1790	$65.50
10	null	null	null	null	null	null	null	5	07/21/1795	$25.50
9	null	null	null	null	null	null	null	6	11/27/1787	$14.40

Full Outer Join



Project Customer Id	customer First Name	customer Last Name	customer Email	customer Address	customer City	customer State	customer Zipcode	orders Order Id	orders Order Date	orders Amount
3	Thomas	Jefferson	tjefferson@u...	931 Thomas J...	Charlottesville	VA	22902	2	03/14/1760	$78.50
2	John	Adams	jadams@usa....	1250 Hancock...	Quincy	MA	02169	3	05/23/1784	$124.00
3	Thomas	Jefferson	tjefferson@u...	931 Thomas J...	Charlottesville	VA	22902	4	09/03/1790	$65.50
10	null	null	null	null	null	null	null	5	07/21/1795	$25.50
9	null	null	null	null	null	null	null	6	11/27/1787	$14.40
5	James	Monroe	jmonroe@usa...	2050 James ...	Charlottesville	VA	22902	null	null	null
4	James	Madison	jmadison@us...	11350 Consti...	Orange	VA	22960	null	null	null

Example:

Table - 1

A	B	C
abc	12234	india
pqr	12234	india
xyz	12234	india

Table -2

X	Y	A
12	ab	xyz
14	cd	abc
15	ef	lmn

Inner Join

A	B	C	X	Y
abc	12234	india	14	cd
xyz	12234	india	12	ab

Full Join

A	B	C	X	Y
xyz	12234	india	12	ab
abc	12234	india	14	cd
lmn	null	null	15	ef
pqr	12234	india	null	null

Left Join

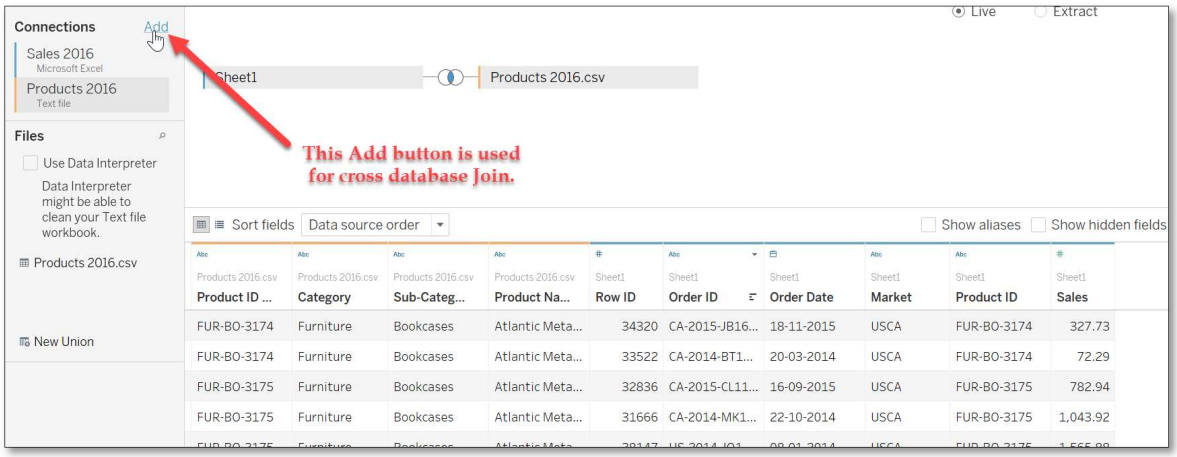
A	B	C	X	Y
abc	12234	india	14	cd
pqr	12234	india	null	null
xyz	12234	india	12	ab

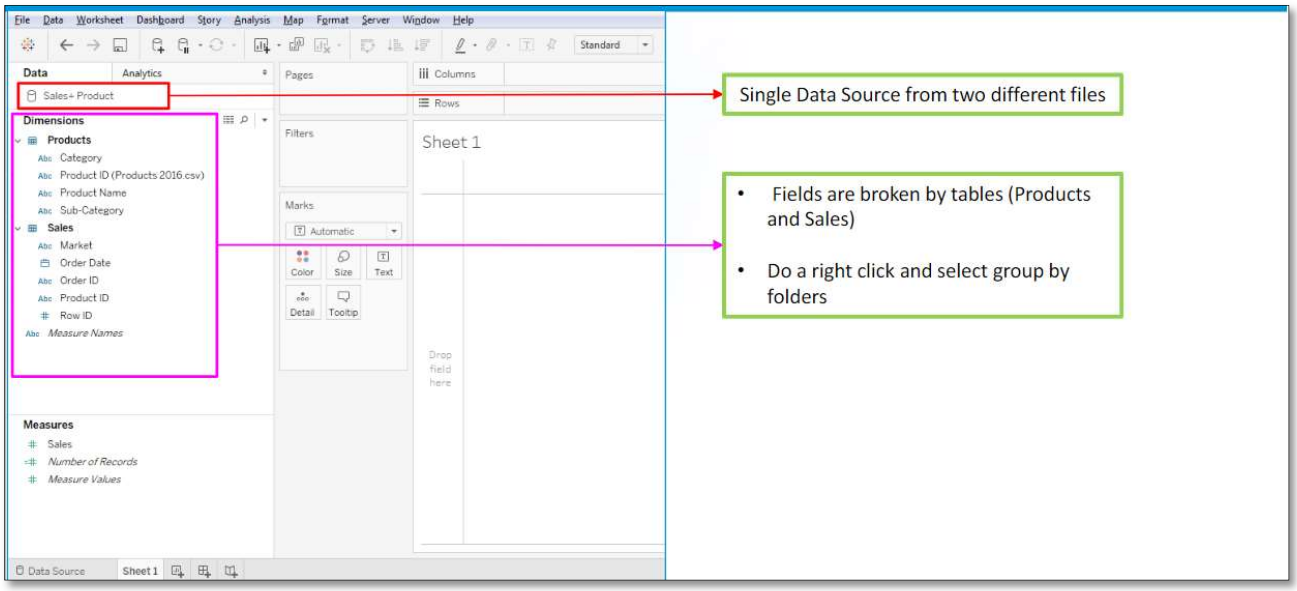
Right Join

A	B	C	X	Y
xyz	12234	india	12	ab
abc	12234	india	14	cd
lmn	null	null	15	ef

5. Cross-database Joins

Many organizations have different data in different systems. Maybe there's financial data in a SQL Server database and product data in Amazon Redshift. The data may live in different environments, but we want to do analysis on everything together.

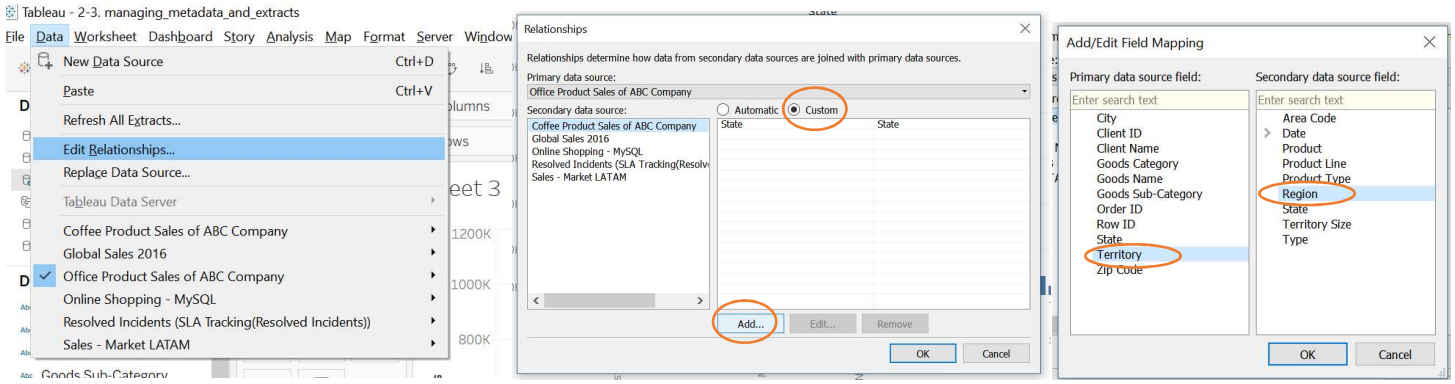




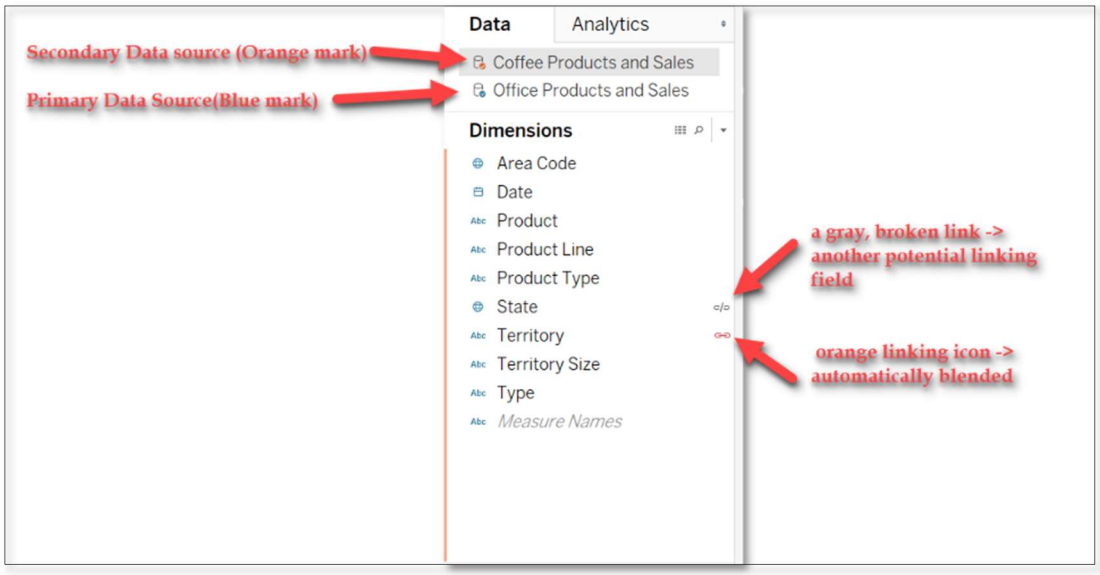
6. Data Blending

Imagine you own two retail chains, Office Products and Coffee Products. You capture your Office Products data in one system and your Coffee Products data in another. Data blending is one way of combining data from multiple data sources into a single view. Instead of joining the data at the row level like a cross-database join, data blending sends separate queries to the separate data sources and aggregates the results to a common level back in Tableau. Data blending requires at least one common field between both data sources.

Example: *State* is common field in our example. *Region* from Office products and *Territory* from Coffee Products refer to same data. We can match it together either by renaming one for the field or by "Data > Edit Relationships" and the adding a custom relationship between them. If they have the same name Tableau can create a relationship between both data sources automatically.



Establishing relationship between Region from Office products and Territory from Coffee Products.



It's important to note that primary and secondary sources are determined on a worksheet-by-worksheet basis and are not maintained globally throughout the workbook. When we are on a new sheet, the data sources within the Data Pane do not have orange and blue check marks to indicate them as primary and secondary. The relationships we established in the previous worksheet are not carried over.

What happens when we swap our data sources and use Coffee Products as the primary source and Office Products as the secondary? We'll drag Coffee Products Sales to the view first, then we'll bring out State. There are now only about half the states on the view compared with our previous example. This is because Coffee Products has fewer states in its data set than Office Products. Tableau is displaying all the states in Coffee Products, and then it will pull in only the relevant information from Office Products. This means that any state in Office Products that is not in Coffee Products will not be displayed in this view.

7. Saving your work!

- **Workbooks (.twb)** – Tableau workbook files have the .twb file extension. Workbooks hold one or more worksheets, plus zero or more dashboards and stories.
- **Bookmarks (.tbm)** – Tableau bookmark files have the .tbm file extension. Bookmarks contain a single worksheet and are an easy way to quickly share your work.
- **Packaged Workbooks (.twbx)** – Tableau packaged workbooks have the .twbx file extension. A packaged workbook is a single zip file that contains a workbook along with any supporting local file data and background images. This format is the best way to package your work for sharing with others who don't have access to the original data.
- **Extract (.hyper or .tde)** – Depending on the version the extract was created in, Tableau extract files can have either the .hyper or .tde file extension. Extract files are a local copy of a subset or entire data set that you can use to share data with others, when you need to work offline, and improve performance.
- **Data Source (.tds)** – Tableau data source files have the .tds file extension. Data source files are shortcuts for quickly connecting to the original data that you use often. Data source files do not contain the actual data but rather the information necessary to connect to the actual data as well as any modifications you've made on top of the actual data such as changing default properties, creating calculated fields, adding groups, and so on.
- **Packaged Data Source (.tdsx)** – Tableau packaged data source files have the .tdsx file extension. A packaged data source is a zip file that contains the data source file (.tds) described above as well as any local file data such as extract files (.hyper or .tde), text files, Excel files, Access files, and local cube files. Use this format to create a single file that you can then share with others who may not have access to the original data stored locally on your computer.